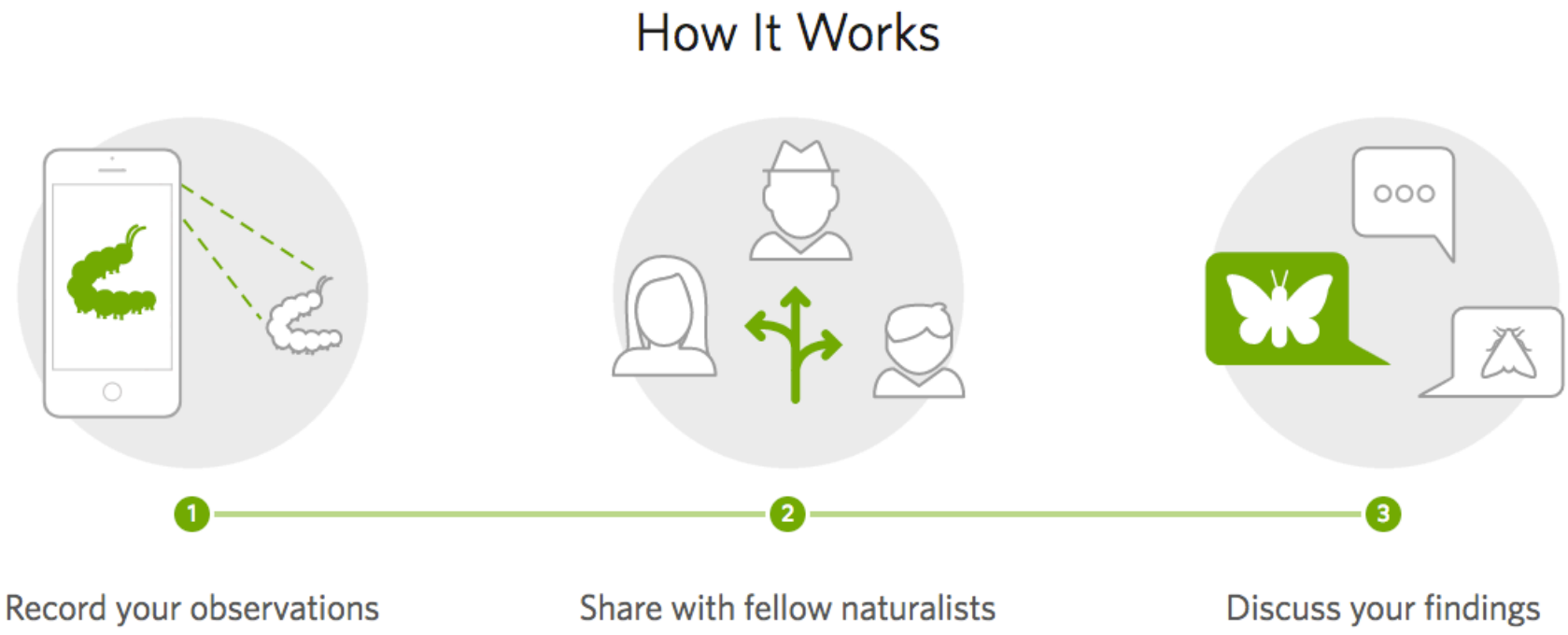
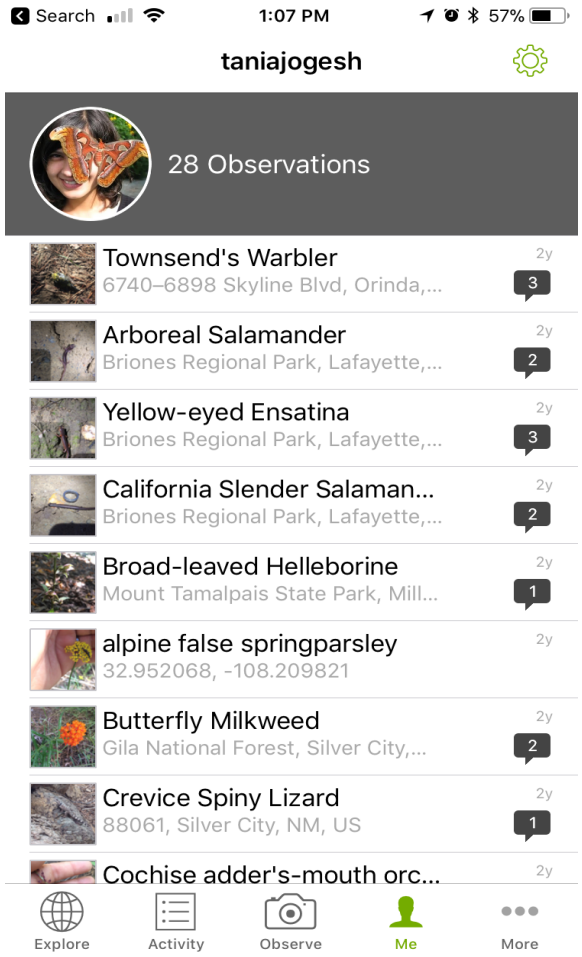
# PREDICTING THE ACCURACY OF CITIZEN COLLECTED BIODIVERSITY DATA

# The iNaturalist App



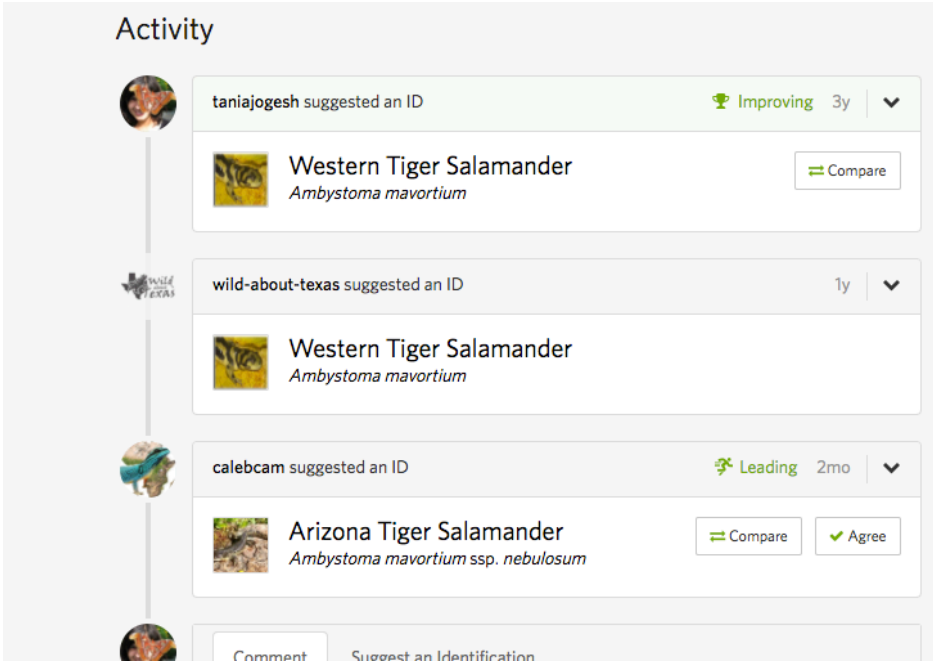
# The iNaturalist Data

- Valuable biodiversity data collected by everyday people
- People share observations of plants and animals with putative identifications
- The community votes on the accuracy of the identifications

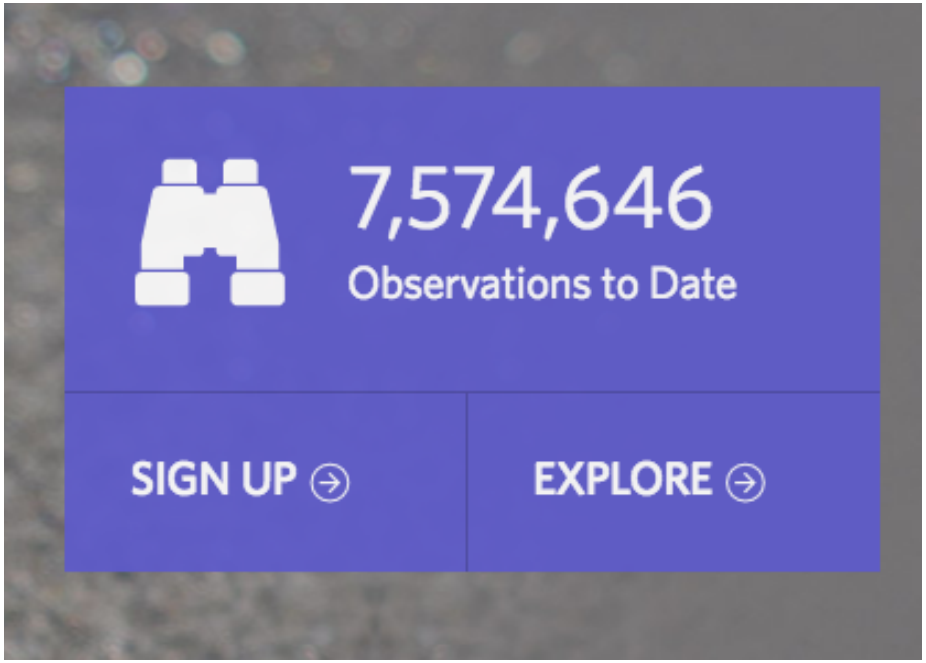


# The Problem

- Correctly identified data provides a treasure trove of information but species are often difficult to identify
- Can we use machine learning to tag observations that are likely to be correct?



# The Data

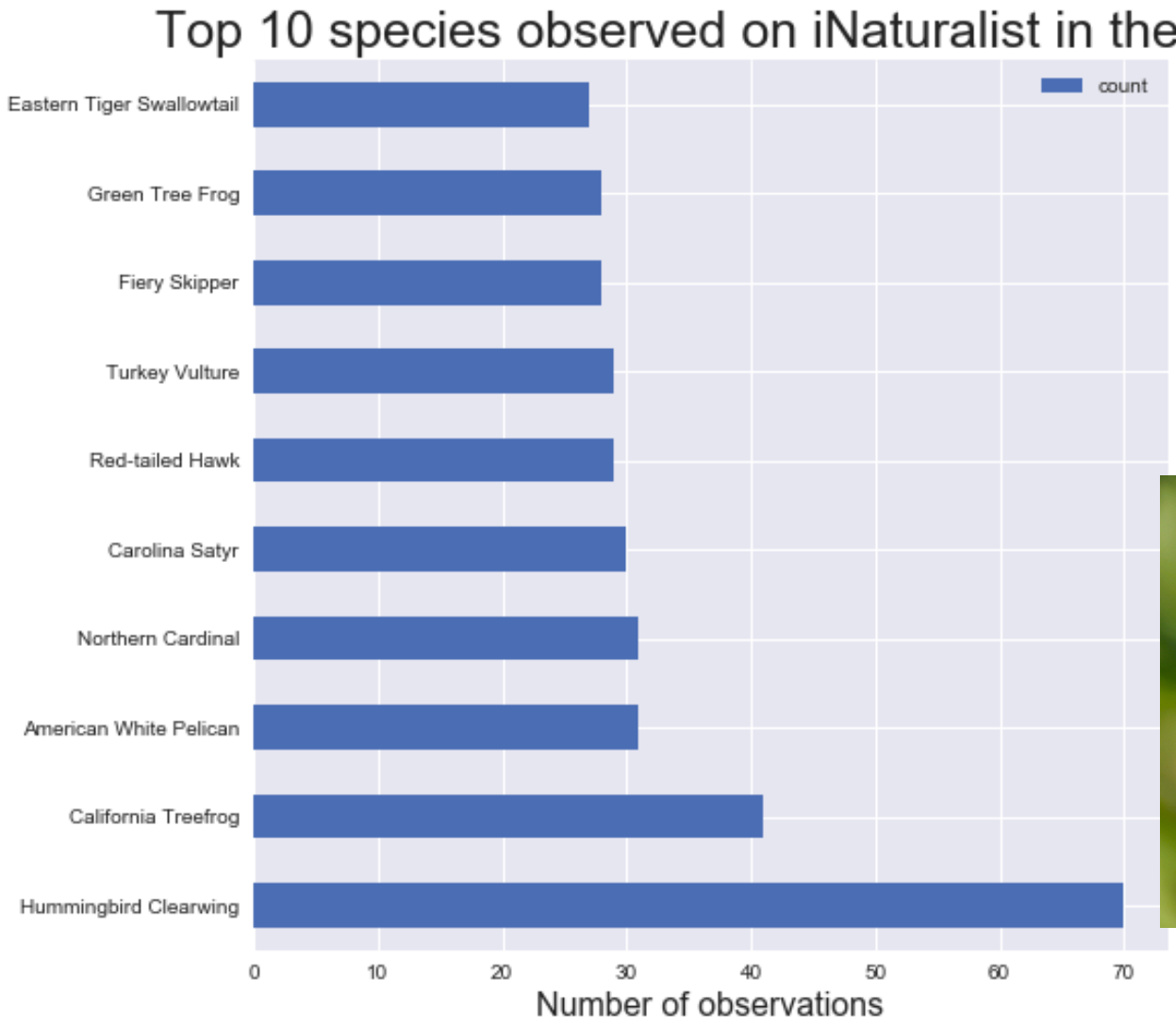



- Queried 10,000 observations using an API
- Limited to data collected between 2016-2017 in North America

# The data

- 450.9 MB of data
- Each observation as individual json file
- Spark Dataframes to read, parse and explore the data

# Top observations

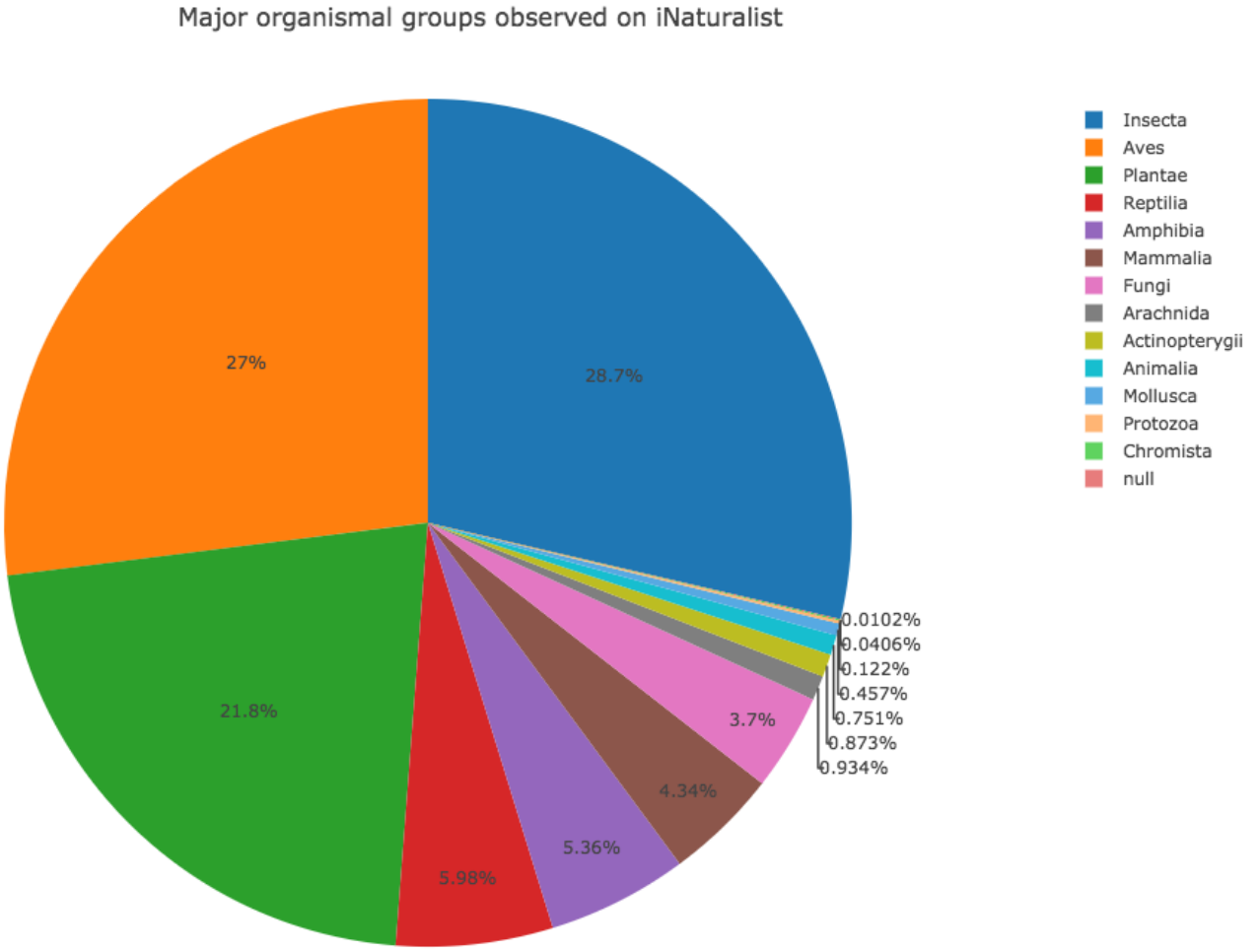



The data suggests that observations on iNaturalist are likely to be biased in favor of species that are more charismatic. Indeed, 80% of the top 10 list are birds and butterflies!

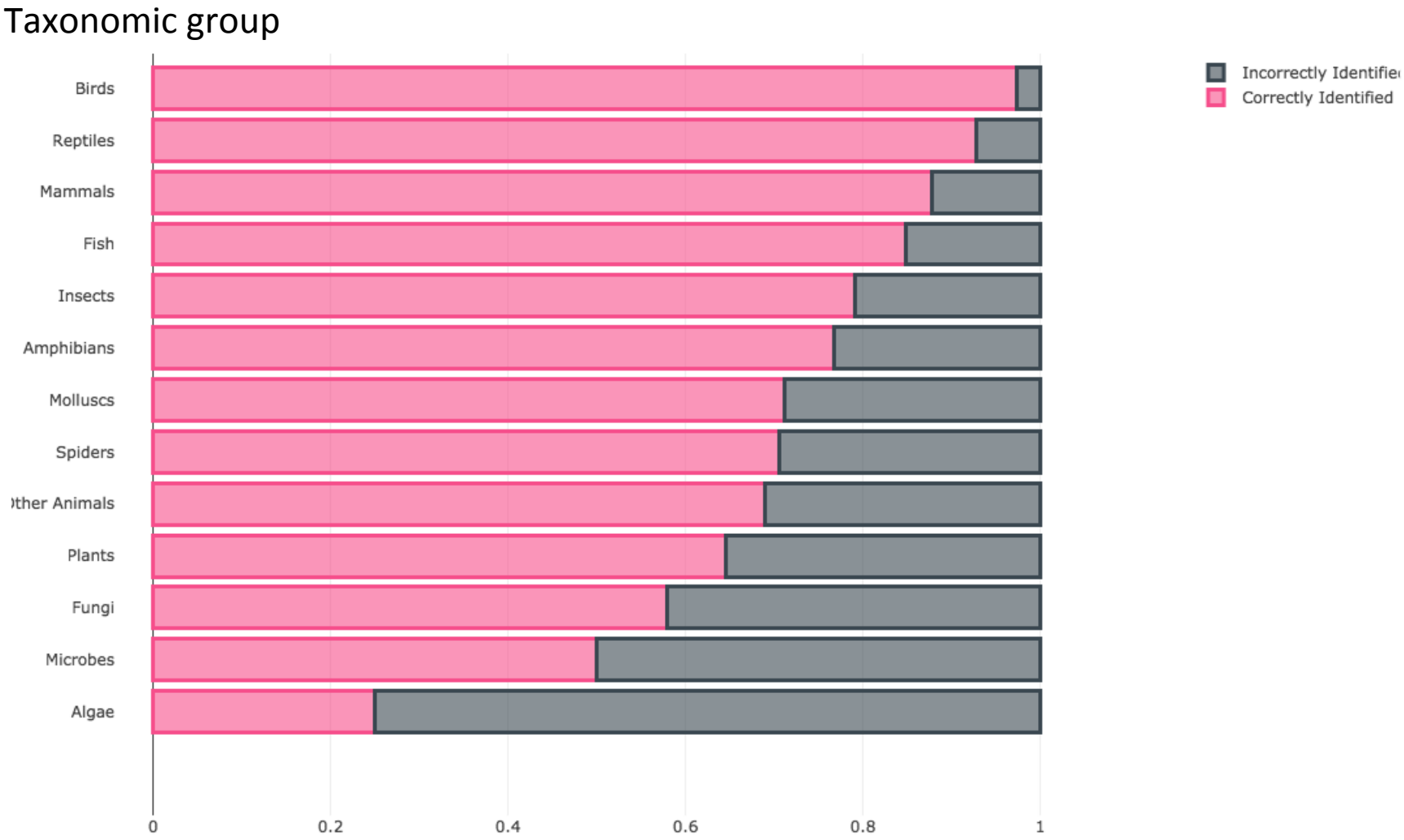
Hummingbird Clearwing

# Major groups observed

Insects, Birds and Plants comprise the bulk of observations on iNaturalist

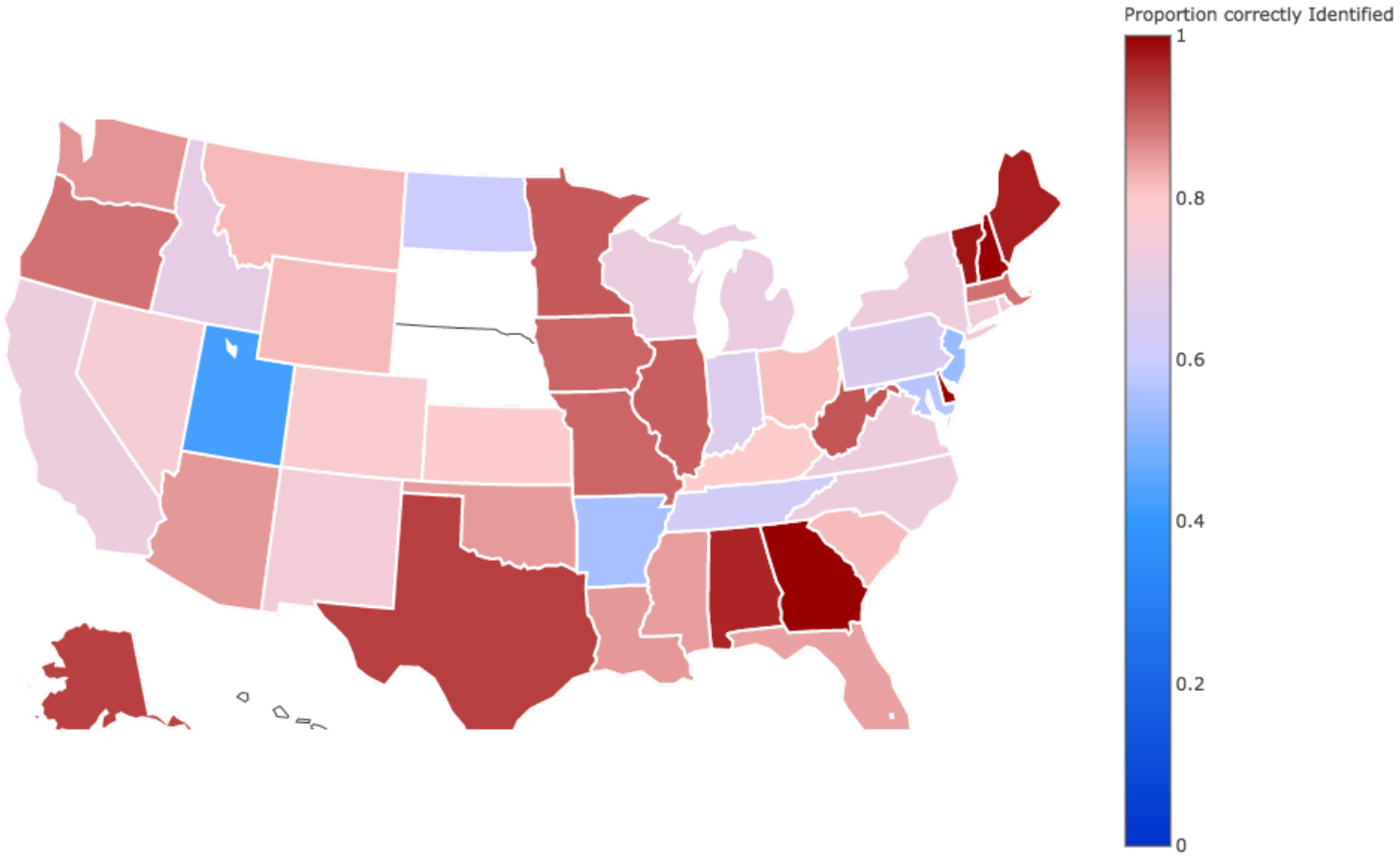


# Features correlated with identification accuracy



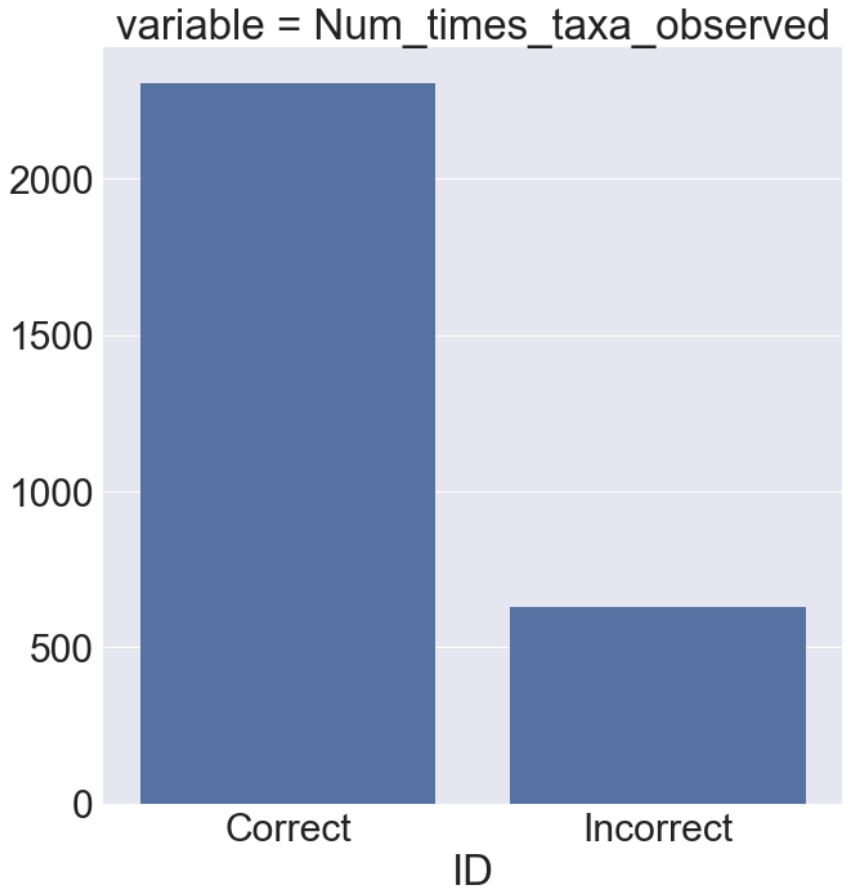
# Features correlated with identification accuracy

Location



# Features that predict identification accuracy

Number of times the species has been observed in the past is a strong predictor of accuracy



# Feature engineering

How to treat categorical variables?

1) encode as numeric labels (not meaningful for linear methods)
2) one hot encoding to create a matrix with each category having a 0 or 1
3) scale features prior to ML

# Machine Learning Models

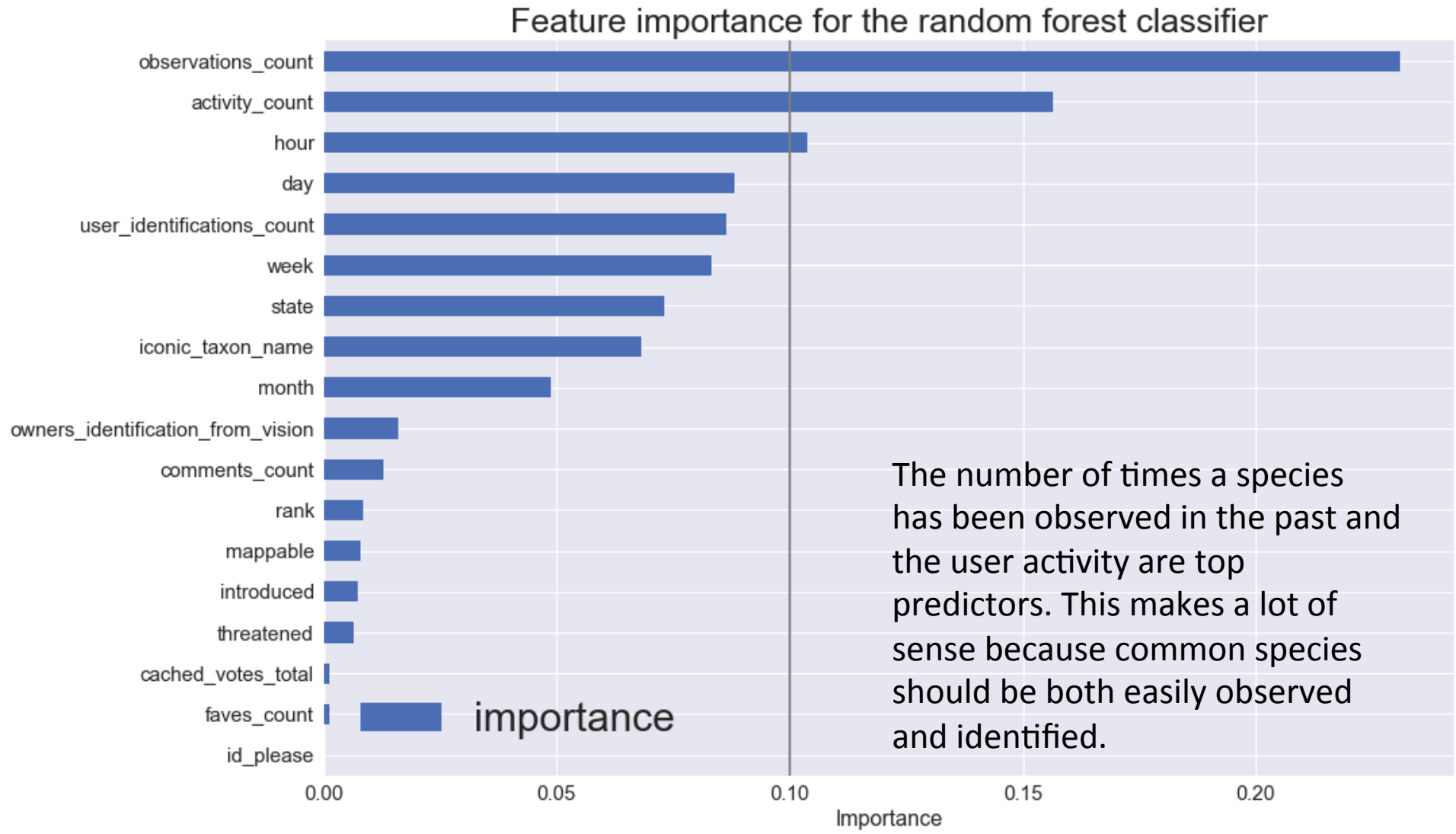
| Model | optimized hyper-parameters | Categorical Features | Train accuracy | Test accuracy | speed |
|---|---|---|---|---|---|
| SVM | C=1, gamma =1 | labeled | 81.2% | 82.1% | 2.03s |
| | | one hot encoded | 87.4% | 83.7% | 21.86s |
| Random forest | default | labeled | 99.7% | 87.6%, | 1.13s |
| | | one hot encoded | 99.7% | 88.2% | 1.99s |
| Gradient Boosting | max depth = 9, learning rate = 0.01, subsample = 0.6, min_leaf=1 | labeled | 99.7% | 87.1% | 33.62s |
| | | one hot encoded | 98.7% | 87.4% | 136.39s |

# Best classifier

| Model | optimized hyper-parameters | Categorical Features | Train accuracy | Test accuracy | speed |
|---|---|---|---|---|---|
| Random forest | | one hot encoded | 99.7% | 88.2%<br>AUC = 0.89 | 1.99s |
| Gradient Boosting | | one hot encoded | 98.7% | 87.4%<br>AUC = 0.87 | 136.39s |

The performance of the Random forest and Gradient Boosting classifier is pretty close. However, the Random Forest wins for being 100x faster with a better AUC score.

# Top features that predict accuracy




The number of times a species has been observed in the past and the user activity are top predictors. This makes a lot of sense because common species should be both easily observed and identified.

# Conclusions

- This classifier can by used by iNaturalist to automatically assign an accuracy rating to new observations.

- Observations with high accuracy can contribute to robust biodiversity data whereas, those predicted to be inaccurate can be tagged for expert identification